Category-rating and magnitude estimation scaling techniques: an empirical comparison

Wegener, Bernd

Dieser Artikel ist eine überarbeitete Version des ZUMA-Arbeitsberichts Nr. 1980/03 "Meßstrukturen und Interskalenrelationen von kategorischen und Magnitude-Skalen" und ersetzt diesen.

The application of a system of measurement structures for category-rating and magnitude scales is tested with regard to numerous sensory and social judgment scales. Results indicate that, on the average, category-rating scales yield interval scales and multimodality matching scales yield logarithmic interval scales. However, marked interindividual differences are detected, pointing to different capabilities subjects have in coping with both methods. It is asked, therefore, how suboptimal scale quality affects the results of parameter estimation in structural equation models in which judgment scales serve as indicator variables. Based on a general psychophysical judgment model it is shown that the measurement theoretical properties of magnitude scales of individual respondents account for large proportions of the variation of estimated coefficients and of goodness of fit. For category-rating scales the effects scale properties have on parameter estimation cannot be determined because of nonhomogeneous judgment functions.

Category-Rating and Magnitude Estimation Scaling Techniques

An Empirical Comparison

BERND WEGENER
Zentrum fuer Umfragen, Methoden, und Analysen
Mannheim, Germany

The two most frequently used direct scaling methods are category-rating and magnitude estimation. The cognitive operations assumed to be basic to each, however, differ and accordingly the two methods are thought to yield different types of scales. This article is aimed at testing the validity of these proposed scale types. In addition, the effects certain scale properties have on the estimation of parameters in structural equations model are studied. I begin by outlining the strategies for coping with both of these problems.

AUTHOR'S NOTE: *I wish to acknowledge the helpful advice and critical comments on earlier versions of this article by Willem E. Saris, Berhard Orth, Peter Schmidt, and anonymous reviewers.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 12 No. 1, August 1983 31-75

In an elaborate category-rating task the subject is provided with a limited number of ordered response categories and is instructed to place that stimulus of a series with the lowest subjective magnitude into the lowest category available and that stimulus with the highest subjective magnitude into the highest. The subsequent stimuli of the series are to be placed into the categories in between, according to the sensational strength the stimuli evoke. Thereby, use of the response categories should be made such that their subjective width is equal. These instructions require the respondent to map subjective differences into response differences; therefore it seems appropriate to assume the category scale to be an interval scale.

In a magnitude estimation task, on the other hand, no response categories are provided; instead, the subject is instructed to choose numbers for the stimuli of a series such that the ratios of the numbers correspond to the "ratios" of the subjective magnitudes of the stimuli of that series. Because magnitude estimations demand the mapping of subjective "ratios" into ratios of number, it seems natural to assume that magnitude estimation scales are ratio scales.

This assumption, however, is justified only if the respondent to a magnitude estimation task follows instructions and forms number ratios in correspondence to subjective "ratios.[1] In a category-rating task, equivalently, an interval scale will result only if the subject forms differences of subjective values and projects these into the equally spaced response categories. Both assumptions cannot be tested directly since, in a direct scaling task, only the category or magnitude scale values as such are given and not the corresponding difference and "ratio" judgments on which the direct judgments are thought to be based internally.

Facing this situation, it is the initial aim of this article to determine empirically the scale properties of category-rating and magnitude estimation scales by testing whether or not judges behave properly, i.e., according to the implicit difference and "ratio" instructions, respectively. A group of measurement theoretical models will be outlined (Orth, 1979, 1982a)—one for the

category-rating scale, one for the magnitude estimation scale, and one for the relationship between the two. These models provide testable axioms and consequences for whether or not individual subjects in the two direct scaling tasks base their judgments on difference and ratio operations, respectively.

The subsequent aim of this article is concerned with practical conclusions. Interest in distinguishing scale types and specifying diverse levels of measurement (Stevens, 1946, 1951) is motivated primarily by considerations with regard to admissible computations on variables of specific characteristics. Suppes and Zinnes (1963) have coined the problem connected herewith "the problem of meaningfulness." In application, of course, meaningfulness is often disregarded. Moreover, with regard to the long-standing "undermeasurement controversy" (Acock and Martin, 1974) it has repeatedly been argued that the level of measurement of a scale is secondary if undermeasurement does not affect the size of relevant indexes compared to "proper" measurement (e.g., Allerbeck, 1978). This inductive argumentation, however, overlooks the fact that suboptimal scale quality may result in misspecified substantial models because the form of relations assumed between variables is dependent on the levels of measurement these variables provide. Therefore, the problem of how the quality of a scale affects the results of substantive models must be coped with by analyzing estimated coefficients and goodness of fit of such models, and not by comparing single indexes.

In addressing this problem this article concentrates on multivariate analyses for which the appropriate level of measurement is the interval scale property. As will be shown in the course of this article, the theory of magnitude estimation provides for a general multivariate judgment model (Saris et al., 1980a; Cross, 1982). For category-rating scales no comparable model is available, since psychophysical and judgment relations for category-rating data are empirically uncertain because their forms are dependent on numerous contextual factors (see, for instance, Parducci, 1982). Lacking a comprehensive theory of categorical judgment and its contextual dependencies, the relationship between scale

quality and results of estimation can be studied, therefore, with regard to magnitude methods only.

Since magnitude estimation theory was originally developed within modern psychophysics—i.e., with regard to the measurement of subjective magnitudes of physical stimuli—the present study aims at comparing results of sensory psychophysical experiments with those of attitude measures. Thus, the two problems to which this article addresses itself are as follows: *Do sensory and attitude scales of direct judgments differ with regard to scale properties, and what are the effects these have on the fitting of substantive models?*

I shall proceed by giving answers to both of these questions by being concerned (1) with the measurement theoretical properties of sensory and social judgment scales and (2) with the effects these measurement properties have on multivariate modeling. In particular, I describe first the category-rating and magnitude estimation measurement structures, reporting the results of testing these structures in the following section. The remaining part of the article is concerned with the effects scale properties of magnitude scales have on the goodness of fit to substantive models. The concept of multivariate psychophysics—sensory and social—is outlined, and following that, the results are described that were found when relating scale properties of individual scales with estimated parameters of the models.

A CATEGORY-MAGNITUDE MODEL

I deal with category-ratings first by outlining a measurement theoretical model for category-rating scales (Orth, 1979, 1982a, 1982b; Orth and Wegener, 1981). Recall that category-rating scales are deprived of any straightforward validation of scale properties. It is, however, assumed that subjects in a category-rating task form differences of subjective values in order to be coherent with instructions that stress that the response categories

should have equal-interval spacing and should be used accordingly. If it is possible, therefore, to gather information about whether or not subjects in fact form internal differences when responding on a category rating scale, the assumption that the resulting scale is an interval scale is testable.

Accordingly, the *category-rating model* consists of two separate parts. In one of these, axioms for difference formation are specified, and these axioms may be tested empirically with regard to *judgments of differences* between pairs of stimuli. It is assumed that these difference judgments map the pairs of stimuli. It is assumed that these difference judgments map the internal difference operations the subject is implicitly instructed to execute during category-rating judgments. As a second component, the category-rating model incorporates a *compatibility condition.* By this condition a link is established between the indirect difference judgments and the direct category-ratings. Only if both are compatible can we be convinced that the difference judgments are indicative for the difference operations that are relevant for the category-rating judgments under study.

Formulating the category-rating model, the following notations are used:

a, b, c, d, ...	=	stimuli a, b, c, d, ...;
ab, bc, ...	=	stimulus pairs ab, bc, ...;
$\gtrsim_D$	=	ordering of subjective differences (a binary relation indicating either "larger" or "equal");
C(a)	=	rating value of stimulus a.

If A is a nonempty set, C a real-valued function on A and $\gtrsim_D$ a binary relation on A X A, the relational structure $\{A, C, \gtrsim_D\}$ is a

category-rating structure if, and only if, for all a, b, c, and d in A the following two conditions hold:

(C1) $[A \times A, \succsim^D]$ is an algebraic difference structure.
(C2) Compatibility between $\succsim^D$ and C holds; i.e.,
$ab \succsim^D cd \leftrightarrow C(a) - C(b) \geq C(c) - C(d)$.

These two assumptions are supplemented by a theorem (1C) and a consequence (2C):

(1C) If $[A, C, \succsim^D]$ is a category-rating structure, the function C is unique up to linear positive transformations; i.e., C is an interval scale.
(2C) $C = a + bD$, where D is the scale to be constructed from the difference judgments, and $a, b \in Re$, $a > 0$ (Re= set of real numbers).

Following the definition of an algebraic difference measurement structure by Krantz et al. (1971: 151ff.), this structure is comprised of five axioms: (1) weak ordering (i.e., transitivity and connectedness), (2) sign reversal, (3) weak monotonicity (or "weak condition on 6-tuples" in Block and Marschak's [1960] terminology), (4) solvability, and (5) an Archimedean axiom. Axioms 4 and 5 are nontestable conditions, which, however, are "almost surely true in the psychophysical context" (Krantz, 1972: 186). In contrast, axioms 1 through 3 of the algebraic difference structure are testable axioms. Axioms 1 and 2 are adequacy requirements for any method by which algebraic difference orderings $\geq$ are obtained empirically (Krantz, 1972: 185). Axiom 3 (weak monotonicity) is thus the main axiom of an algebraic difference structure that remains to be tested empirically.

However, following Orth (1979, 1982a) this study does not test weak monotonicity, but the so-called quadruple condition (Block and Marschak, 1960). As Block and Marschak showed (1960: 112ff.) this condition implies weak monotonicity, though weak monotonicity does not imply the quadruple condition. Thus, the quadruple condition is a stronger necessary condition for an algebraic difference struc-

ture to exist, and in empirical testing it is preferable to weak monotonicity for a conservative estimate of the amount of violations. For the ordering $\gtrsim_D$ of the category rating model, the quadruple condition has the following form:

$$(C1.Q) \quad ab \gtrsim_D cd \rightarrow ac \gtrsim_D bd$$

Inasmuch as only C1.Q is tested, an application of the category-rating model is tested—not the complete model as such.

The compatibility condition C2 states that the rank order of difference judgments should coincide with the rank order of the numerical differences between the corresponding category-rating values if difference and category-rating judgments are to be compatible. If this condition is fulfilled empirically, the category-rating scale is an interval scale, given that the difference judgments do not violate the axioms of the algebraic-difference structure (C1). This statement is true due to Theorem 1C, which is a straightforward consequence of the representation and uniqueness theorem of an algebraic difference structure (Krantz et al., 1971: 151ff); the respective proof is given by Orth (1979: 68).

The consequence 2C follows from Theorem 1C: Given the validity of C1 and C2, both scales C and D are interval scales for the same set of stimuli, and as such they are linearly related because interval scales are defined by the class of positive linear transformations as permissible transformations.

Turning to the *magnitude estimation model*, the situation looks quite similar to that of the category-rating model. In magnitude estimation, however, the respondents are instructed to make implicit ratio judgments instead of difference judgments; therefore, the magnitude estimation model provides conditions by which this assumption can be tested empirically. Making use of the following notations:

a, b, c, d, ...	= stimuli a, b, c, d, ...;
ab, bc, ...	= stimulus pairs ab, bc, ...;
$\gtrsim_R$	= ordering of subjective ratios;
M(a)	= magnitude value of stimulus a,

the assumptions of the magnitude estimation model have the form (Orth, 1979, 1982a, 1982b; Orth and Wegener, 1981):

If A is a nonempty set, M a positive-valued real function on A, and $\gtrsim_R$ a binary relation on A X A, the relational structure $\{A, M, \gtrsim_R\}$ is a magnitude-estimation structure if, and only if, for all a, b, c, and d in A the following two conditions hold:

(M1) $\{A \times A, \gtrsim_R\}$ is an algebraic difference structure.
(M2) Compatibility between $\gtrsim_R$ and M holds, i.e., $ab \gtrsim_R cd \leftrightarrow M(a)/M(b) \geq M(c)/M(d)$.

Again, there are a theorem (1M) and a consequence (2M):

(1M) If $\{A, M, \gtrsim_R\}$ is a magnitude-estimation structure, the function M is unique up to power transformations, i.e., M is a logarithmic interval scale.
(2M) $M = aR^b$, where R is the scale to be constructed from the ratio judgments, and $a, b \in Re$, $a, b > 0$.

The magnitude-estimation model is in close agreement with the category-rating model, but it applies to ratios instead of "differences." The correspondence of the two models is due to the well-known fact that a difference representation may be represented by several different numerical scales, one of these being a ratio representation of exponentially transformed differences (Krantz et al., 1971: 152). However, instead of yielding an interval scale, a ratio representation of differences is unique up to any power transformation with positive constants; thus, the resulting scale is a logarithmic interval scale, as stated in theorem 1M. The proof of that theorem is straightforward (Orth, 1979: 77), and 2M is a testable consequence of 1M, given that conditions M1 and M2 are fulfilled empirically.

By the same arguments given above, the axiom of weak monotonicity of the algebraic difference structure for "ratios" may be replaced by the stronger quadruple condition of Block and Marschak (1960), which has the form

(M1.Q) $ab \geq_R cd \rightarrow ac \geq_R bd$

within the context of the magnitude estimation model. Finally, it should be noted that the magnitude estimation model may be extended to apply to cross-modality matching and is, as such, an ingredient part of "relation theory," as proposed by Krantz (1972) and Shepard (1978).

Note that the magnitude estimation model proposes that magnitude estimation scales are logarithmic interval scales, not ratio scales. This is opposed to Stevens's claim (1975) that magnitude estimation scales, based on psychological "ratios", have ratio scale level being unique up to similarity transformations (multiplication by a positive constant). It has been shown, however, that Stevens's claim is justified if the *cross-modality matching system* is taken into account, which involves several magnitude scales on several reaction modalities (besides numerical magnitude estimates, e.g. sound, line, or force of handgrip production). The cross-modality system assumes that these scales are related to each other by power functions. Inasmuch as this is true, it is evident that the scales will be ratio scales if *one* of the interscale functions is fixed to a constant exponent (Krantz et al., 1971: 165). By this restriction the magnitude scales are free to vary only by multiplication with a constant, and this operation defines ratio scales.

However, one may doubt the usefulness of nonrelative exponents of the cross-modality matching system (Shepard, 1978). Empirically the exponents can be determined only in their relation to each other, and not by absolute size (Cross, 1974, 1982). How then can the claim that magnitude scales are ratio scales be validated other than by fixing one of the exponents arbitrarily?

The approach suggested here may be outlined as follows: Suppose that for one and the same set of stimuli the category-rating and the magnitude estimation models hold for a category-rating and magnitude estimation scale, respectively. The difference representation of the category-rating scale C will thus be an interval scale and the ratio representation of the magnitude estimation scale M will be a logarithmic interval scale. Therefore, a

$(C + \kappa)$ and a_2M^b are also difference and ratio representations for differences and ratios, respectively. From this it can be seen that if a_1, a_2, b and κ can be chosen such that $a_1(C + \kappa) = a_2M^b$, then both transformed scales are ratio scales, because any solution of the proposed interscale relation will determine uniquely the constants b and κ but not a_1 and a_2, the quotient of which may be specified only. Thus, similarity transformations on both sides of the interscale equation are performed, and this type of transformation characterizes ratio scales.

If a joint ratio scale along these lines should result, the following assumptions of a joint *category-rating-magnitude-estimation model* must be met empirically with regard to a category-rating scale and C and a magnitude estimation scale M (Krantz et al., 1971: 153).

If A is a nonempty set, C a real-valued and M a positive-valued real function on A, and if $\succeq_D$ and $\succeq_R$ are two binary relations on A X A, the relational structure $\{A, C, M, \succeq_D, \succeq_R\}$ is a category-rating magnitude-estimation structure if, and only if, for all a, b, c, d, e, and f in A, the following four conditions hold:

(CM1) The category-rating model (C1, C2, 1C, and 2C) is valid for C.
(CM2) The magnitude estimation model (M1, M2, 1M, and 2M) is valid for M.
(CM3) $\succeq_D \succeq_R$ are distinct relations.
(CM4) The interlocking condition holds, i.e.,
  (i) $ab \succeq_R aa \leftrightarrow ab \succeq_D aa$;
  (ii) $ad \sim_R be \sim_R cf \rightarrow [ab \succeq_D bc \leftrightarrow de \succeq_D ef]$.

This set of conditions yields the following two theorems:

(1CM) $a_1(C+\kappa) = a_2M^b$ or $(C+\kappa) = aM^b$; κ, a, b ϵ Re and a, b > 0.
(2CM) $(C+\kappa)$ and aM^b is a ratio scale.

The assumption CM3 is a necessary condition if the results should be obtained. If the orderings $\succeq_D$ and $\succeq_R$ are *not* distinct the

relation between the category-rating and the magnitude estimation scale will be a logarithmic relation (Orth, 1979: 82ff). Formally, if CM4 is omitted and instead of CM3,

(CM3*) $\gtrsim_D$ and $\gtrsim_R$ are not distinct relations,

then

(1CM*) $C = a + b \log M$; $a, b \in Re$ and $b > 0$.

Note that nondistinctiveness of the difference and the ratio relations indicates that subjects are unable to distinguish between "differences" and "ratios." That this is so empirically has been conjectured by Torgerson (1960, 1961) and has been provided with further empirical evidence, e.g. by Birnbaum (1982). The category-rating-magnitude-estimation model, however, proposes a *two representations theory*, according to which respondents are capable of executing distinct difference and ratio operations. This point of view is also taken, for instance, by Marks (1974) and Krantz (1972).

The interlocking condition (CM4), introduced by Krantz et al. (1971), specifies how difference and ratio relations must be interrelated qualitatively in order to yield a common scale of two distinct representations. This condition is difficult to test empirically, however, and it is not considered in the subsequent parts of this article. Theorems 1CM and 2CM are direct consequences of the results obtained by Krantz et al. (1971: 158-163); see also Orth (1979: 85-87).

This concludes the specification of the three axiomatic models for category-rating and magnitude estimation scales. I turn now to empirical data.

PHYSICAL STIMULI

In this section the results of testing the three axiomatic models with regard to sensory judgment scales are reported. For sensory scales the models have been tested before in one instance (Orth,

1982a); no use of multimodality matching, however, was made in that study.

PROCEDURE

Eighteen social science students participated in the experiment. Their primary task was to evaluate the length of single straight lines as well as the relationship between two lines presented simultaneously. Thus, direct estimates and indirect, stimulus-pair judgments were assessed.

The *stimuli* that were presented to the subjects in individual experimental sessions were nine lines of a geometrically spaced series ranging from 1 to 31 cm in objective length. The lines were presented to the subjects horizontally on a monitor screen; they were shown in random order, each line appearing twice for every direct and indirect judgment task. In single stimulus presentations the lines were shown one at a time and a pair of lines was presented for the indirect judgments. Subjects were seated in front of the screen at about 2.5 m distance, such that their eyes leveled with the projected lines.

In front of them the subjects had a keyboard device for making category and numerical magnitude *responses* as well as a turnable knob for adjusting sound production responses to be listened to over earphones. Sound was centered at $f_o = 3125$ Hz and regulated by a sone taper potentiometer; the impulses appeared in a one second on-off rhythm, and their intensity could be adjusted from 36.5 decibels (db) to 100 db in steps of 1/2 db. The knob for adjusting the tones was automatically attenuated each time subjects had chosen a response intensity.

Equipped with these devices subjects received training in making magnitude judgments with numbers and sounds. When they reported that they understood the tasks, they were asked to give the following series of judgments, differing in order:

(1) numerical magnitude estimation (ME) of single lines; no response standard provided;

(2) sound production responses (SP) of single lines (without standard);
(3) category-ratings of single lines on a numerical nine-point scale labeled "small" and "large" at its endpoints;
(4) pair-difference judgments of all possible pairs of lines; subjects were to express the differences on a numerical nine-point scale from "least different" to "most different";
(5) pair-ratio judgments of the pairs with numbers; instructions called for subjects to input that number (real number or decimal) by which multiplying the smaller of the two lines would give the longer one;
(6) pair-ratio judgments of the pairs by sound ratios; in this case the subjects were instructed to equate the weakest producible sound with the smaller line of each pair and to choose that sound intensity for responding that seemed x times as strong to them as the longer line exceeded the smaller one.

RESULTS

Quadruple Conditions

Recall that the three axiomatic models, first of all, call for the testing of the algebraic-difference structures with a difference representation for the pair-difference judgments and a ratio representation for the pair-ratio estimates. The crucial axiom regarding this is the weak monotonicity condition. In this study the somewhat stronger quadruple condition (C1.Q and M1.Q) is tested. The mean percentages of individual violations of the quadruple conditions are given in Table 1, the left panel referring to the difference judgments, the middle panel to the ratio judgments with numbers, and the right panel to the ratio judgments with sounds. The mean percentages of violations refer to those cases in which the antecedents of the quadruple axiom hold but the conclusion fails.

Obviously, more violations occur for difference compared to ratio tasks. But interindividual variations of results are quite considerable. Violations of the quadruple axiom for sound ratios, for instance, varies from no violations at all to over 12% for different subjects. Before drawing any conclusions, therefore,

TABLE 1
Mean Percentages of Violations of Quadruple Conditions for Difference and Ratio Judgments

	DIFF	RATIO-NU	RATIO-SO
	QUAD	QUAD	QUAD
Median	9.28%	1.12%	2.63%
Mean	9.13%	2.02%	4.06%
S.D.	3.01	1.63	3.30

regarding whether or not the measurement structures are fulfilled and to what extent this is so, interindividual differences will have to be considered with regard to all the testable conditions of the three models. I will return to this issue later in this article.

Compatibility Conditions

The next assumption to be tested with regard to the category-rating and the magnitude estimation models is compatibility (C2 and M2). Compatibility is given if the rank order of the indirect judgments coincides with the rank order of the differences or ratios of the corresponding direct judgments. Table 2 exhibits the results of computing the rank correlations; Kendall's τ corrected for ties was used. Each cell of Table 2 gives mean coefficients and, in parentheses, the standard deviations for the eighteen subjects. Abbreviations are as follows: DIFF-LL, differences of physical line lengths; DIFF, difference judgments; DIFF-CAT, differences of category-ratings; RAT-LL, ratios of physical line lengths; RATIO-NU, "ratio" judgments with numbers; RATIO-SO, "ratio" judgments with sound; RAT-ME, ratios of magnitude estimates; RAT-SP, ratios of sound productions; RAT-MAG, ratios of the combined magnitude scale constructed by geometrically averaging magnitude estimates and sound production responses for each individual.

TABLE 2
Mean Rank Correlation Coefficients for "Differences" and "Ratios"

	1	2	4	5	6	7
1 DIFF-LL						
2 DIFF	.722 (.121)					
3 DIFF-CAT	.792 (.075)	.764 (.085)				
4 RAT-LL	---	---				
5 RATIO-NU	---	---	.616 (.140)			
6 RATIO-SO	---	---	.672 (.183)	.503 (.120)		
7 RAT-ME	---	---	.776 (.132)	.517 (.157)	.634 (.191)	
8 RAT-SP	---	---	.700 (.137)	.449 (.211)	.560 (.181)	.691 (.107)
9 RAT-MAG	---	---	.856 (.097)	.530 (.209)	.636 (.193)	---

NOTE: See text for explanation of labels. Standard deviations in parentheses.

The results most interesting to us are the rank correlations between DIFF/DIFF-CAT (= .764), between RATIO-NU/RAT-ME (= .517), and between RATIO-SO/RAT-SP (= .560). All three values are substantial but lower than expected. Note, however, that the rank correlations between the involved responses and the differences and ratios of the *physical* line lengths are quite low also; one would not expect to find rank correlations between responses that exceed these correlations to any notable degree. Note also, that again there is considerable variation between individuals. As will be shown later on, there is evidence that this variation is systematic in nature and that we have to consider the fact that subjects have different capabilities and are differently equipped to cope with difference and "ratio" tasks.

Linearity and Power Relations

Still being concerned with average results, I turn to the consequences of the models. If the category-rating scales are interval scales, they should be related linearily to the one-dimensional scales of differences such that $C = a + bD$. Scale D may be constructed by smallest space analysis of the difference judgments (Schneider, 1982). Conversely, if the magnitude estimation scales are logarithmic interval scales, these scales should form power function relationships with the one-dimensional scales of the ratio judgments that result from smallest space analyses: $M = aR^b$. Fitting the respective functions the mean correlation coefficients of Table 3 result, Thereby the power relations are estimated by linearizations; it should be noted, however, that a MINISSA-scale of ratio judgments is based on a difference representation; therefore, the R-scales are transformed exponentially and the functions $\log M = a + bR$ have been fitted instead of $\log M = a + b \log R$ (Compare Schnieder et al., 1974; Orth, 1982a: 365).

In Table 3 the following abbreviations are used: C is the category-rating scale and D the one-dimensional scale of difference judgments; ME and SP are the magnitude estimation and sound production scales, respectively; R_{NU} is the MDS-scale of numerical "ratio" judgments and R_{SO} the MDS-scale of "ratio" judgments with sound intensities; MAG, finally, is the combined magnitude scale constructed from geometric averaging of responses in both modalities. As can be seen from Table 3 the goodness-of-fit values for the functions are quite high.

Interscale Relations

The testable theorem of the joint category-rating-magnitude-estimation model is the prediction of the relationship betwen category-rating and magnitude estimation scales. This relation should be an additive power relation ($C + \kappa = aM^b$) under the assumption that the difference operation $\gtrsim_D$ and the "ratio" operation $\gtrsim_R$ are distinct (1CM), and it should be a logarithmic relation ($C = a + b \log M$) under the assumption that both operations are identical (1CM*).

TABLE 3
Average Correlation Coefficients for the Linear and Power Relations Between Direct and MDS Scales

	median R	mean R	S.D.
$C = a + bD$	.977	.969	.023
$ME = a R^{b}_{NU}$	.971	.954	.081
$ME = a R^{b}_{SO}$	.971	.950	.078
$SP = a R^{b}_{NU}$	.947	.931	.046
$SP = a R^{b}_{SO}$	.948	.939	.035
$MAG = a R^{b}_{NU}$	.969	.964	.025
$MAG = a R^{b}_{SO}$	.971	.966	.022

According to the joint model the distinctiveness of the difference operation from the "ratio" operation can be tested by comparing the rank order of differences of the category-rating scale values with that of the corresponding ratios of the magnitude scale. We find that Kendall's τ is roughly .40 in mean value if the differences calculated on the category-rating scales are compared with the ratios of the magnitude scales in the number and sound modalities. In general we thus expect that the power relation between the direct category-rating and magnitude scales will yield a better fit than the logarithmic relation. As can be seen from Table 4, this prediction is corroborated even though the goodness of fit expressed as mean correlation coefficients for the (linearized) additive power functions exceeds that for the logarithmic functions only slightly. The parameters of all interscale relations were estimated by the iterative procedure of Wegener and Kirschner (1981). These authors as well as Wegener (1982b) gave evidence for the general applicability of the power model in several thousands of scale comparisons in various domains.

TABLE 4
Mean Correlation Coefficients for Power and Logarithmic Interscale Relations

Category-ratings as functions of:	R(pow)	b	R(log)
ME	.962 (.053)	.490 (.301)	.937 (.030)
SP	.945 (.035)	.137 (.167)	.931 (.030)
MAG	.970 (.017)	.285 (.225)	.955 (.020)

NOTE: b is the mean exponent of the power relation fit; standard deviations in parentheses.

The second column of Table 4 may explain why superiority of the power model is small in the present case. In this column the mean values for the exponent b of the power function relationship between category-rating and magnitude scales are listed. As can be seen, these estimates do not only differ for the two magnitude modalities, the large amount of individual variation is also noteworthy. This points to the fact that a number of subjects yield a fairly small interscale exponent. In these cases a logarithmic interscale relation might fit the data equally well. (See Wegener [1982b] for an explanation of the form of the interscale relation as a function of the ranges of the scales related.) Therefore, even though in general the power form describes the interscale relationship more adequately, a logarithmic relation may capture the relation of both types of scales equally well for some individuals. In terms of a general statement then we again have to consider interindividual differences before conclusions with regard to the interscale model can be drawn. Before this is done, however, we turn to the scaling of attitudes in order to compare the validity of the three axiomatic models in the sensory domain with that in the social domain.

SOCIAL STIMULI

PROCEDURE

Forty-six subjects of two age groups (16-35 years, 36 years and above) and three different levels of school education were quota-sampled and were interviewed in their homes by professional interviewers. Primarily, they were asked to express their opinions toward sixteen occupations. The occupations were symbolized by respective titles and were chosen to cover the full range of the International Occupational Prestige Scale by Treiman (1977), female factory worker being the lowest and physician the highest profession of the set. The sixteen occupational titles had also been subject to a previous study in which occupational prestige was measured by a cross-section survey of over 2000 respondents (being representative for the West German population of 16 years and older) by category and magnitude methods (Wegener, 1982b). In the present study the subjective prestige of the 16 occupations was measured also. In addition, subjects had to scale the social importance and the average standard of living associated wtih the professions. However, only the prestige measures were designed for testing scale properties. For this purpose the 46 respondents were required to execute the following tasks:

(1) Category-rating of prestige on numerical 9-point and a 20-point rating scales; two rating scales were used because different rating scales often yield different results (Parducci, 1982).
(2) Numerical magnitude estimations and magnitude "line production" of prestige; in line production (LP) respondents are asked to express relative ratios of subjective intensities by drawing horizontal lines differing in length (see Wegener, 1980, 1982b, for details of procedure).
(3) Difference judgments of prestige differences with regard to all possible pairs of 9 of the 16 occupations; a 20-point numerical rating scale was used for the subjects to indicate subjective differences; based on previous studies the 9 occupational titles

had been selected such that their prestige values were approximately evenly distributed over the total range of the set of 16 occupations; all respondents judged the same 9 occupations (pairwise).

(4) Ratio judgments of prestige ratios with regard to the same pairs of occupations; respondents were asked to give numerical estimates of subjective ratios.

Subjects were handed a booklet of format 30 × 21 cm in which to make their responses. They were briefly trained in magnitude estimation and line production by means of sizes of circles and seriousness of offenses that were to be evaluated with the two reaction modalities. The procedure had previously been tested and applied in a large series of experiments and field studies (Wegener, 1978, 1979, 1980, 1982b; Beck et al., 1979).

RESULTS

Quadruple Conditions

In Table 5 the mean percentages of violations of the quadruple conditions of the difference and ratio judgments are given. A comparison with Table 1 shows that these results are similar to those for physical stimuli. On the average, more violations are encountered for the difference than for the ratio tasks but, in absolute terms, there are slightly less violations for difference and slightly more violations for ratio judgments in comparison to the sensory data sets.

Compatibility Conditions

The degree of compatibility of difference judgments with category-ratings and of ratio judgments with magnitude responses is expressed in Table 6. Mean τ values for difference judgments (DIFF) and the vectors of differences of the 9- and 20-point category scales, respectively (DIFF-C_{09} and DIFF-C_{20}), and those for the ratio judgments (RATIO) and the ratios of the (combined) magnitude scores (RAT-MAG) are given. In addition, the rank correlations between the differences (ratios) of the

TABLE 5
Mean Percentages of Violations of the Quadruple Axiom

	DIFF	RATIO-NU
	QUAD	QUAD
Median	7.47%	4.51%
Mean	8.06%	4.76%
S.D.	4.74	2.93

Mean German prestige scale from the survey of over 2000 respondents (Wegener, 1982b) are shown (DIFF-2000 and RAT-2000). The latter values constitute the upper expected bound for the rank correlations between the involved responses of the present study.

Linearity and Power Relations

Table 7 shows means of correlation coefficients for fitting the category scales (both the 9-point scales [C_{09}] and the 20-point scales [C_{20}]) to the scales constructed from the difference (D) judgments (via one-dimensional MDS-solutions), as well as the results of fitting the combined magnitude scale of prestige (MAG) to the one-dimensional scale of ratios (R). The values are slightly lower than the corresponding ones of Table 3 from the sensory domain, but they still express a high degree of fit of the functions.

Interscale Relations

Results with regard to estimating the interscale relations and testing whether 1CM or 1CM* is valid are given in Table 8. As in Table 4 the correlation R(pow) and the exponent b for the additive power model and the correlation R(log) for the logarithmical

TABLE 6
Mean Rank Correlations Between Pair Judgments and Differences and Ratios of Category-Ratings and Combined Magnitude Scores, Respectively

	1	2	5	6
1 DIFF-2000				
2 DIFF	.614 (.091)			
3 DIFF-C_{09}	.651 (.140)	.630 (.147)		
4 DIFF-C_{20}	---	.550 (.201)		
5 RAT-2000	---	---		
6 RATIO	---	---	.586 (.079)	
7 RAT-MAG	---	---	.546 (.163)	.561 (.168)

NOTE: Standard deviations in parentheses.

interscale model are listed as mean values. Both the 9- and the 20-point category-rating scales were fitted to the combined magnitude scale MAG.

Mean values of b and the variation over individuals suggest that the power functions for a number of subjects have rather small exponents, and for these subjects a logarithmic function may fit the data equally well. In spite of the good mean fit of the interscale relations it is mandatory therefore to inspect interindividual variation closely before a general statement about the validity of the joint model can be made.

TABLE 7
Goodness of Fit of Functions Relating Direct Judgment Scales with One-Dimensional Scales of Differences and Ratios

	median R	mean R	S.D.
$C_{09} = a + bD$	.926	.873	.193
$C_{20} = a + bD$	.905	.824	.281
$MAG = aR^b$	.904	.879	.111

Summarizing the results of the testing of the three axiomatic models for social stimuli, it may be concluded that no striking difference in results was obtained in comparison to the study of sensory attributes. All goodness-of-fit index tested for are roughly identical in size in spite of the differences in stimulus kind, experimental set-up, and response modalities. This stability is especially noteworthy with regard to violations of the quadruple condition (Tables 1 and 5), which is the main necessary condition for empirical tests of the respective algebraic difference structures. In both studies, the pair-stimulus judgments are in closer agreement with the measurement theoretical models for "ratio" than for difference tasks. Both studies, however, exhibit a large degree of variation of individual results, and the question to be answered next is whether or not this variation is systematic in nature. If it is random the conclusion to be drawn is that the proposed models have no general applicability. However, if the interindividual differences reveal a consistent pattern with respect to specific groups of subjects the models may be said to be valid but conditional on individual characteristics.

INTERINDIVIDUAL DIFFERENCES

On these lines it can be proposed that subjects differ in their abilities according to a simple fourfold table resulting from a

TABLE 8
Mean Correlation and Exponent b for Fit of the Power Interscale Relations and Mean Correlation for Logarithmic Interscale Relations

Category scales as function of MAG	R(pow)	b	R(log)
09-point	.848 (.100)	.364 (.470)	.838 (.104)
20-point	.861 (.101)	.637 (.609)	.844 (.101)

NOTE: Standard deviations in parentheses.

dichotomy of "difference capability" and a dichotomy of "ratio capability." This gives four types of individuals, those who are good at difference and ratio tasks, those who are good only at ratio tasks and not at difference formation, those who are good at differences and not good at ratios, and finally the individuals not good at either task. If this typology in fact mirrors personal differences in judgment consistently, we expect to find a group of subjects having high score indexes on all the tasks executed, another group of subjects scoring high only on tasks in which ratio formation was asked for, a third group with high scores on differences-related tasks, and finally another group of subjects with detrimental results in all instances.

This psychological hypothesis—central to the controversy of "differences" versus "ratios" in direct scaling—was tested straightforwardly by means of a factor analysis of the individual results of the complete tests of the axiomatic models over all subjects. Separate analyses were executed for each of the two studies. In both cases two-factor solutions result, explaining 58% of the total variance in the sensory and 71% in the social stimulus study. Loadings on the varimax rotated factors clearly reveal a difference and a ratio factor for both data sets.[2] Furthermore, dichotomizing the factor scores of both factors of the two sets at their

median places each subject in one of the cells of the fourfold table. In Table 9 the mean values of the relevant indexes for the four groups of the prestige study are displayed, DIFF+/RATIO+ representing the group of positive "difference" and "ratio" capabilities, DIFF+/RATIO- the group of positive "difference" and negative "ratio" capabilities, and so on.

This classification displays the consistent pattern of variation of individual indexes, which is revealed by the two-factor solution: Subjects with positive "difference capabilities" produce better indexes with regard to difference related tasks—low amount of violations of the quadruple axiom, high correlations with regard to compatibility and fit of functions—than subjects with negative "difference capabilities"; the same is true for "ratio type" subjects with regard to the respective indexes. Also, a very similar pattern is found for the study of sensory stimuli. Therefore, we conclude that the three models do not fit the judgments of all subjects in the same way; rather, we have to distinguish different types of capable individuals who *consistently* follow either "difference instructions" or "ratio instructions" or both, while a fourth group does not seem to be affected by either type of instructions at all. This result is of importance for the attempt to develop a general theory of direct judgment, and it sheds new light on the ratio-difference controversy in psychological measurement (Birnbaum, 1982).

The finding is equally important for applied research and substantive theory construction. In scaling subjective phenomena, not all respondents produce the same quality of scales, and the individual levels of measurement vary with different abilities for performing direct scaling tasks. It is therefore objectionable to treat subjective scales with identical means of analyses, even if the scales come from one and the same study. From a pragmatic point of view, this conclusion, however, has impact only if it can be demonstrated that the quality of a scale affects the outcome of substantive analyses. The remaining part of this article, therefore, is devoted to showing that level of measurement does in fact matter and that interindividual differences in scale quality must be taken into consideration when analyzing direct scaling data.

TABLE 9
Group Mean Values of Difference and Ratio "Capabilities"

	DIFF + RATIO+	DIFF + RATIO-	DIFF - RATIO+	DIFF - RATIO-	TOTAL
N of subjects	14	9	9	14	46
Quad. DIFF violations	5.03%	5.36%	11.26%	10.56%	8.06%
Quad. RATIO violations	2.71%	4.37%	4.35%	7.33%	4.76%
Compatibility:					
C_{09} - DIFF (Tau)	.73	.62	.59	.52	.61
C_{20} - DIFF (Tau)	.69	.57	.54	.39	.54
M - RATIO (Tau)	.65	.47	.64	.42	.54
$C_{09} = a + bD$ (R)	.941	.924	.827	.806	.873
$C_{20} = a + bD$ (R)	.917	.917	.756	.705	.824
$M = aR^b$ (R)	.945	.817	.904	.838	.880
$C_{09} + \kappa = aM^b$ (R)	.925	.784	.889	.788	.848
$C_{20} + \kappa = aM^b$ (R)	.931	.768	.918	.821	.861

MULTIVARIATE PSYCHOPHYSICS

MAGNITUDE MEASUREMENT

The model to study the effects that different levels of measurement have on substantive parameters is outlined in Figure 1, representing the *general psychophysical judgment model* (Saris et al., 1980a; Cross, 1974, 1982). The model is applicable whenever magnitude measurements of subjective phenomena are involved irrespective of whether the stimuli are sets of physical or social entities. In order to see why this is so, a brief glance at psychophysical theory is called for.

Psychophysics is the study of the relationships between physical intensities and the strength of perceptual impressions associated with these. In literally thousands of experiments it has been confirmed that this relation has a power function form if the physical entities are measured in energy values and if strength of

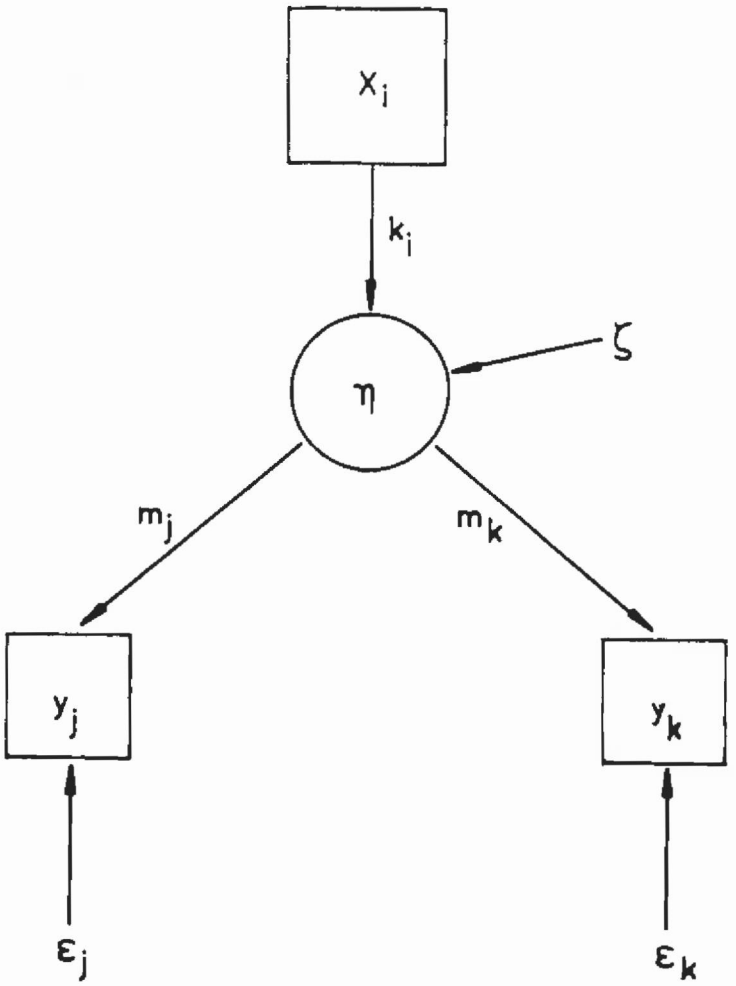


Figure 1: General Psychophysical Judgment Model

sensation is assessed by direct magnitude estimates (Stevens, 1975). Strictly speaking, Stevens's power law does not apply to the relation between stimuli and sensations but to the relation between stimuli and (magnitude) responses. If proportionality,

however, is proposed between numerical magnitude estimation responses and the subjective continuum—as was done, for instance, by Stevens (1975)—the cross-modality matching system yields a reformulation of the power law consisting of two separate power relations: one between stimuli and sensations and one between sensations and observable responses. Moreover, introducing multiplicative measurement errors, the psychophysical power law may then be expressed with two fundamental equations:

$$\Psi = S_i^{k_i} e_\Psi \qquad [1]$$

$$R_j = \Psi^{m_j} e_j \qquad [2]$$

In these formulae S_i symbolizes the vector of physical intensities of modality i, R_j is the vector of magnitude responses on reaction modality j, and Ψ indicates the vector of associated sensations. All three vectors have the same number of elements, depending on how many distinctive stimuli are presented. Both equations, of course, may be expressed in linear form based on the following definitions:

$$y_j = \log R_j;\ \eta = \log\Psi;\ \epsilon_j = \log e_j;\ \zeta = \log e_\Psi;\ x_i = \log S_i$$

Instead of equations 1 and 2, we then have

$$\eta = k_i x_i + \zeta \qquad [1']$$

or

$$y_j = m_j\eta + \epsilon_j \text{ or } \underset{\sim}{y} = \underset{\sim}{m}\eta + \underset{\sim}{\epsilon} \qquad [2']$$

if several response modalities are used as in sensory-modality matching. For two response modalities Figure 1 schematizes both structural equations. Hauser and Goldberger (1971) discussed the properties of equivalent MIMIC models. Note that in the application proposed here, however, the involved linear vectors have a

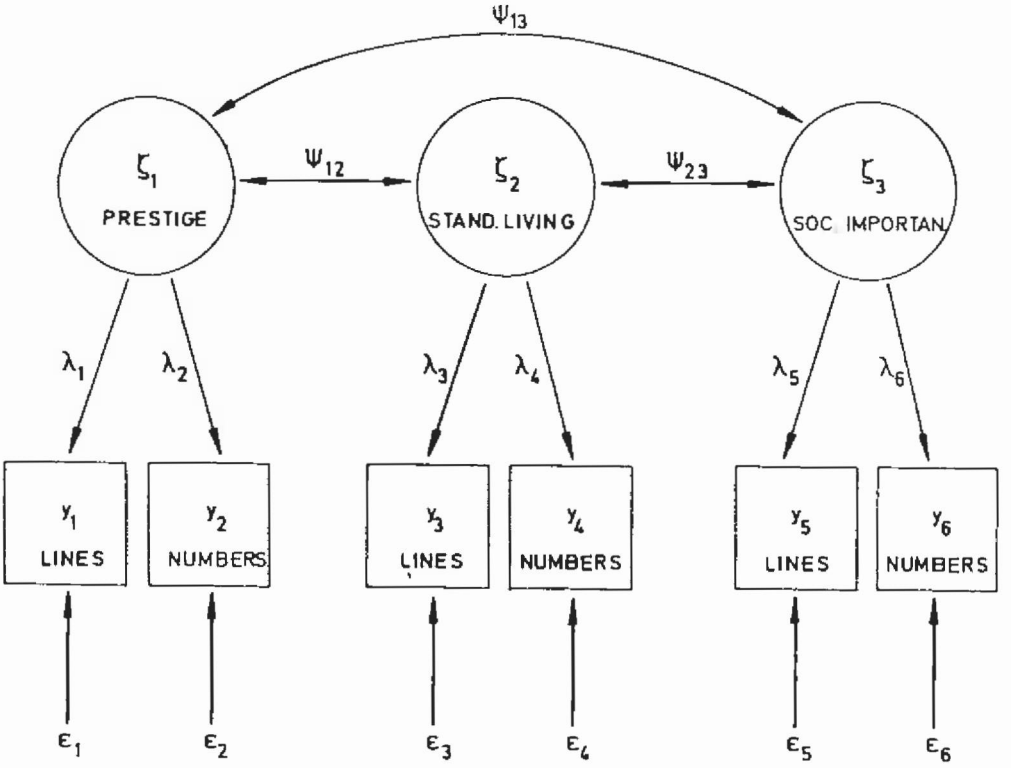
length identical to the number of *stimuli* presented, thus enabling separate analyses for every respondent for whom stimulus-response measures are available.

Turning to social psychophysics, we have to do without exogenous variables, leaving us with just one structural equation or sets of structural equations, namely equation 2′. Because of conventional convenience, we change the notation of equation 2 to equation 3:

$$y_j = \lambda_j \zeta + \epsilon_j \text{ or } \underset{\sim}{y} = \underset{\sim}{\lambda}_y \zeta + \underset{\sim}{\epsilon} \qquad [3]$$

Equation 3 describes a Congeneric Test model (Joereskog, 1971, 1974). For three constructs and six indicators an example is given in Figure 2. It should be remembered that response values enter this model in logarithmic form because of the assumed power relations. If magnitude scales are logarithmic interval scales, as proposed by the axiomatic magnitude estimation model, the transformed indicators are treated as having interval scale properties.

In its simple form the model of Figure 2 assumes correlations between three latent variables of occupational cognition: prestige of occupations, the standard of living associated with occupations, and the social importance ascribed to the occupations. The concepts of these three variables have received some attention within the functionalistic theory of stratification since the influential work of Davis and Moore (1945). The respective empirical studies (e.g. Reiss, 1961; Hodge et al., 1964, 1966; Treiman, 1977) have made use of averaged category-rating scales of occupational attitudes, and they yield stable correlations between different aspects of judged occupations and a remarkable intersocietal and intrasocietal agreement. It has been argued that both results are methodological artifacts (Coxon and Jones, 1978; Wegener, 1979) due to the crude scaling methods and aggregation strategies used. In the study of occupational cognition reported here, however, bimodal sensory-modality matching with numbers and lines was used for the assessment of the indicator variables, assuming


Ψ13
ζ1
PRESTIGE
Ψ12
ζ2
STAND. LIVING
Ψ23
ζ3
SOC. IMPORTAN.
λ1
λ2
λ3
λ4
λ5
λ6
y1
LINES
y2
NUMBERS
y3
LINES
y4
NUMBERS
y5
LINES
y6
NUMBERS
ε1
ε2
ε3
ε4
ε5
ε6

that, compared to category-ratings, a more veridical measurement is achieved.

CATEGORY MEASUREMENT

In contrast to magnitude scales, category-rating scales cannot be used as indicator variables within the general psychophysical judgment model or its extensions. The model specifications assume power relations between all involved observable and latent variables, and this assumption rests on well-corroborated empirical evidence for magnitude measurement only. With regard to category-rating scales, conversely, the forms of the stimulus-response functions, the psychophysical functions (e.g., k_l in Figure 1), and the judgment functions (e.g., m_j and m_k in Figure 1) are inherently unstable and dependent on numerous contextual conditions (see, for instance, Parducci, 1982). It is for this reason that, in contrast to Stevensian psychophysics, research on category measurement capitalizes on contextual effects, in addition to the response procedures, in order to seek lawful relationships between both (compare Birnbaum, 1982: 407; Wegener, 1982a: 29-33). In spite of impressive progress in that direction (Birnbaum, 1982: 427-449), what Luce and Galanter had deplored almost 20 years ago is still true—namely, that a "sophisticated theory of categorical judgment . . . which defines a scale of sensation that is invariant under the various experimental manipulations" is not available (Luce and Galanter, 1963: 268). Given the diversity of sources for possible variations of the category scale any type of monotone function interrelating stimuli, true scores, and responses is conceivable. In comparison with magnitude measurement, therefore, category measurement is lacking a rationale for choosing a homogeneous type of function on which to base multivariate judgment models.

Facing this discrepancy of the levels of theorizing with regard to the two kinds of methods, it is not feasible to assess the effects the scale properties of category-rating scales have on the results of substantive analyses with regard to individual subjects. If, however, for individual respondents, category scales can be fitted to

Figure 2: Congeneric Test Model for Occupational Evaluation of Prestige, Standard of Living, and Social Importance, with Numbers and Lines

magnitude scales satisfactorily—in accordance with the interscale relation of ICM—it may be assumed that the respective psychophysical and judgment functions, as well as scale properties, are identical for both types of scales. Therefore, the effects scale quality has on substantive estimation may be inferred to be identical for both in these instances. We are left without such information, however, in all remaining cases.

Considering this limitation, in the section to follow the results of fitting the outlined judgment models to the magnitude data are presented, and the relationship between measurement theoretical properties of the judgmental indicators and estimated model parameters is analyzed. Both the sensory and the social data sets are studied.

LEVEL OF MEASUREMENT AND MODEL FIT

PHYSICAL STIMULI

For each of the 18 individuals of the line-length study, a MIMIC model (Figure 1) with two indicators (numerical magnitude estimation and sound production) and one exogenous variable (line-length stimuli) was fitted using Joereskog and Soerbom's (1978) Analysis of Linear Structural Relationships (LISREL). Note that in this application of the method the number of observations equals the number of stimuli, and this number is quite small in the present case. According to Monte Carlo experiments (Boomsma, 1982), however, the maximum likelihood estimates that are approximately unbiased for large samples seem to be unbiased for small samples as well. In computation, however, small sample sizes increase the risk of no convergence and also the probability of improper solutions giving negative estimated unique variances. Keeping these difficulties in mind, LISREL can be used for the error-free estimation of psychophysical and judgment functions if the number of stimuli is small.[3] Moreover, the method allows for tests of psychophysical judgment models for each individual separately.

For the data sets of the physical stimuli, estimations of the model parameters are possible if coefficients m_j for the magnitude estimates are fixed at 1.0 (leaving the variance of η unstandardized). No tests of model fit are feasible, however, since degrees of freedom (df) = 0. Based on the individual s covariance matrices, the mean estimated coefficent m_k for sound production is 2.054 (S.D. = 1.081) with mean k_i = .889 (S.D. = .266). Because of the logarithmic transformations of the observed variables, k_i, and m_j and m_k are exponents of the psychophysical input and output functions, respectively. ($k_i \cdot m_k$ is an error-free estimate of the exponent of the stimulus-response function for sound production). For the 18 subjects its mean value is 1.636 (S.D. = .413), which is in close agreement with the size of exponents reported for pooled data in psychophysical experiments (Stevens and Galanter, 1957). Note that the present computation averages *parameters* of individuals whereas the usual psychophysical procedure averages response *scores*. In Table 10 mean factor loadings of standardized solutions of the input and output variables are given, and it can be seen that these values are highly satisfactory.

After the simple MIMIC models were fitted to the data sets of the 18 subjects the following procedure was chosen to test whether or not scale quality has an effect on the estimated parameters. For all subjects the relevant indexes of the measurement theoretical analyses were treated as independent variables in multiple regression analyses, with the estimated individual parameters of the model fitting as dependent variables. Thus, the main independent variables were (a) percentages of violations of the quadruple axioms, (b) rank correlations with regard to the compatibility conditions, (c) goodness-of-fit measures with regard to the power relations between the magnitude scales and the scales of "ratios," and (d) goodness-of-fit measures of the power interscale relations.[4] These independent variables were used to predict the indicators' factor loadings of the standardized solutions and the standard errors of estimates as well as the deviations of the product exponent $k_i \cdot M_k$ from the "theoretical" value of 1.6, which is to be expected when line length is matched by loudness (see Stevens and Guirao, 1964). In Table 11 the multiple correla-

TABLE 10
Mean Factor Loadings of Standardized Solutions of Input and Output Variables and Standard Deviations

Factor loadings of standardized solutions:	Mean	S.D.
numbers	.972	.067
sounds	.939	.063
lines	.984	.034

tion coefficients resulting from these regression analyses are given. As can be seen, the predictions are very good; that is, there is an impressive influence on the model parameters by the scale qualities the individuals produce. Inasmuch as the indicator scales are (log-) interval scales, factor loadings of the standardized solutions increase and standard errors of estimates decrease and, in addition, the estimations of exponents agree more closely with psychophysical predictions.

SOCIAL STIMULI

Are these effects also observable with regard to social judgment scales? In order to answer this question the data sets of the occupational evaluation study were used and the model of Figure 2 was fitted to the magnitude scales of judged prestige, standard of living, and social importance for each of the 46 subjects of that study. In accordance with psychophysical assumptions the λ-coefficients for lines were fixed at 1.0 (the variances of the three construct variables were left unstandardized). Since this model yielded poor fit for quite a number of individuals, five additional versions were tested alternatively in order to account for the possibility of correlated error variances of identical indicator modalities. Including the original model as model 1 these were the following:

TABLE 11
Multiple Correlations of Regressing Measurement Theoretical Indexes on Model Parameters

Dependent variables	Multiple R
Factor loadings of standardized solutions of numbers	.985
Factor loadings of standardized solutions of sounds	.968
Standard error numbers	.934
Standard error sounds	.993
$\lvert 1.6 - (k_i \cdot m_k) \rvert$	.986

Model 1: No correlations occurred between error variances of the six indicators (as in Figure 2).

Model 2: Correlated error variances between the three line indicators are assumed only.

Model 3: Correlated error variances between the three number indicators are assumed only.

Model 4: All error variances of identical modalities correlate with each other, but the correlations between lines are constrained to the same value (numbers free).

Model 5: All error variances of identical modalities correlate with each other, but the correlations between number indicators are constrained (lines free).

Model 6: Correlations between error variances of line indicators are constrained, and those between number indicators are also constrained.

The results of these 6×46 analyses of the individual's covariance matrices are shown in Table 12. In its left panel Table 12 gives the summed χ^2 values and the summed degrees of freedom over all subjects for each of the six tested models. The evaluation of results is based on the differences of χ^2 values and degrees of freedom of sequentially compared pairs of models (Bentler and Bonett, 1980). On these lines model 4 is the best-fitting average model since the improvement in χ^2 is—relative to the corresponding degrees of freedom and in comparison to the other five

TABLE 12
Goodness of Fit of Six Hierarchical Models

	ALL SUBJECTS			SUBJECTS' BEST MODELS		
Model	$\sum \chi^2$	$\sum df$	N	$\sum \chi^2$	$\sum df$	N
1	435.95	276	46	21.80	36	6
2	234.65	138	46	14.64	15	5
3	174.00	138	46	12.88	15	5
4	91.20	92	46	13.67	32	16
5	104.87	92	46	5.76	14	7
6	232.77	184	46	14.17	28	7

NOTE: χ^2 and df summed over all subjects (left panel) and summed over subject's best models only (right panel).

models—greatest for that model. This need not be so for each individual, however. If one chooses that model for which the fit is best *relative to each subject* (terming it the "best model" for that subject) the right panel of Table 12 results.[5] We see that 16 subjects have model 4 as their best model, and the remaining subjects yield the best fit with one of the other five models. Considering only the best models, the summed χ^2 and df values indicate again that model 4 has the best fit; but for those individuals for which one of the other models is the best model, individual parameter estimation should be based on those models.

Model 4 is that model in which the correlations between error variances of the line indicators are constrained to the same value; the respective correlations between number indicators are left unconstrained. Because of this restriction it is quite plausible that model 4 turns out to be the best-fitting model for all subjects as well as with regard to the individually best models. As will be remembered from an earlier section of this article, respondents of the occupational cognition study were to make their responses in a booklet; it is conceivable that the format of that booklet (with a paper width of 30 cm) influenced the line responses to the three scaling tasks for each individual in the same manner. This influ-

TABLE 13
Mean Factor Loadings of Standardized Solutions of Indicators of Subjects' Best Models

Factor loadings of standardized solutions of:	Mean	S.D.
y_1 Prestige with lines	.932	.157
y_2 Prestige with numbers	.922	.158
y_3 Stand. Living with lines	.977	.180
y_4 Stand. Living with numbers	.910	.133
y_5 Soc. Importance with lines	.874	.318
y_6 Soc. Importance with numbers	.969	.291

ence seems to be responsible for the identity of correlations of error variances between the individuals' line vectors.

Estimated from subjects' best models, the mean factor loadings of the standardized solutions of Table 13 result. The factor loadings of the six indicators are quite high compared to the values to which attitude researchers are accustomed. For similar results, see Saris et al. (1980b) and Saris (1981).[6]

Next, the question of how level of measurement influences the goodness of fit and individual parameters of the models is considered. The procedure here parallels the analysis of the line-length study: The measurement theoretical indexes that relate to the magnitude scaling of occupational prestige (violations of axioms, compatibility, power relation fit, and goodness of fit of the power interscale relation) enter multiple regression analyses as independent variables in order to predict several model parameters. Among these are the probability level for rejecting the fit of the specific model, the factor loadings of the standardized solutions of indicators, and the standard errors of estimate with regard to coefficients. In line with psychophysical reasoning we also predict the amount of deviation from 1.0 of the individual λ-coefficients for numbers; the exponent (i.e., the λ-coefficient) for numerical magnitude estimation should be unity since the corresponding

TABLE 14
Multiple Correlations of Regressing Measurement Theoretical Indexes on Model Parameters of Subjects' Best Models

Dependent variables	All S.s.	Model 4
Probability level	.531	.609
Factor loadings of standardized solutions of lines	.585	.731
Factor loadings of standardized solutions of numbers	.440	.547
Stand. error lines	.509	.619
Stand. error numbers	.610	.758
$\|1.0 - \lambda_2\|$	.606	.762

exponent for line production has been fixed to 1.0 in the model specification.

The results of these multiple regression analyses are shown in Table 14 by giving the respective multiple correlation coefficients. The left of the two columns consists of the coefficients based on all subjects' best models; the right column gives the results of analyses based on subjects having model 4 as their best model only. Even though the multiple correlation coefficients are somewhat lower than those of the line length study (Table 11), the results convincingly demonstrate the effects that scale properties have on parameters estimation in attitudinal models.

CONCLUSIONS

In summarizing, the main results of this research may be stated thus:

(1) In direct scaling it seems that category-rating scales yield interval scales and magnitude estimation scales yield logarithmic interval scales. This is so, however, only if one considers different types of judges. Not all subjects are equally capable of coping with the two methods; rather, we have to distinguish "category type" subjects and "magnitude type" subjects. For those subjects, however, who conform to both groups and who produce the required level of

measurement from both kinds of tasks category-rating scales and magnitude scales may be transformed into ratio scales and, moreover, the relation between the two kinds of scales is of an additive power form. These findings are in agreement with Krantz's (1972) and Shepard's (1978) "relation theory" as well as with the category-magnitude models of Orth (1982a). In addition, this article validates the claim that the results apply regardless of whether sensory or social judgments are studied.

(2) When based on the general psychophysical judgment model magnitude scales give highly satisfactory results in multivariate modeling. These results parallel the findings of Saris et al. (1980a) and of Saris (1981). There is evidence, however, for a strong relationship between the goodness of fit and estimated coefficients of these multivariate models and the levels of measurement of the indicator variables. Again, these results apply to psychophysical scaling and to the scaling of attitudes as well.

(3) For category-rating scales no parallel assessment of the effects scale properties have on substantive analyses is feasible, due to the unsystematic nature of psychophysical and judgment functions with regard to these scales. Clearly, research on magnitude scaling techniques has progressed farther, compared with that on categorical judgment, and this situation is reflected in the possibility of formulating individual judgment models for magnitude measurement and of validating these empirically while judgmental processes related to categorical measurement must still be explored.

The most important conclusion to be drawn from these findings is that any attempt to classify "types of subjective scales" by types of direct scaling methods has ambiguous results. From the data sets analyzed in this study, one may conclude that magnitude scaling procedures do only *tend* to produce logarithmic interval scales and category-rating methods do only *tend* to produce interval scales, but by no means can we be certain that this is so in general. Subjects seem to have an a priori predilection for either one of the two types of methods and for the "ratio" or "difference" logic characteristic for these methods, respectively. In psychological scaling research the distinction between ratio-based and difference- (or similarity) based scales has proved to be a very

useful one (Marks, 1974), and this distinction may in fact be deduced from axiomatic assumption (Wegener, 1982a). Unlike indirectly assessed scales, however, which have validity by definition (Luce and Edwards, 1958), it is uncertain which of the two types of scales that we encounter with direct scaling methods. Difference in instructions does not guarantee differences in results. At this point we can only speculate about whether the resistance of certain individuals to judge according to specific instructions is caused by innate factors, obstrusiveness, or habit.

The other conclusion to be drawn from the finding of interindividual variation in types of produced scales is concerned with the effects this variation has on data analysis. In the present form, this problem has not been considered in the scaling literature before, since a certain scaling method used for assessing indicators is usually believed to yield a specific level of measure-ment homogeneously for all involved subjects or aggregate pseudosubjects. The present research demonstrates—for magnitude measurement—that this belief is unwarranted and that, moreover, suboptimality in scale quality distorts the quality of substantive model estimation. On these grounds, it is conjectured here that the analyses of direct sensory or social judgment scales, or systems of such scales, lead to artifactual results unless the levels of measurement appropriate for these analyses are secured. As long as we are deprived of the knowledge of what are the characteristics a person must have in order to respond to a direct scaling task properly (i.e., to produce the required level of measurement), individual tests for his or her scale properties should be provided by following the procedures applied in this article.

In this respect it is important to note that in multivariate analyses of sensory or social response scales, insufficient levels of measurement may amount to errors of specification, detectable only by determining the measurement theoretical properties of these scales. Linearity in structural equations is based on interval scales, or on logarithmic interval scales inasmuch as the general psychophysical judgment model is applied. If linearity is assumed, contrary to this requirement, the respective model is misspecified, since its relations may not be linear at all. Such a misspecification

may yield poor fit of the model, or high measurement error, or both. In either case it is unknown whether these are substantial results or whether they are due to inappropriate properties of the scales under study.

These words of caution allude to both category-rating and magnitude estimation scales. This study, however, was able to demonstrate the detrimental effects of suboptimal scale properties only with regard to the latter. Due to the uncertain nature of internal judgmental relations and contextual dependencies of the category scale, it is not possible to establish a general judgment model for category scales that parallels the magnitude model and with reference to which equivalent effects could be studied. Reflecting on the question of which mode of direct scaling should be used for specific purposes, one should therefore consider this asymmetry in the level of theorizing with regard to the two methods. Since the results of the measurement theoretical analyses put forth in this article are in some respects discouraging (in that not all respondents seem to be equally capable of handling either or both methods adequately), it is of special importance to control for resulting defects when using judgment scales in sociological exploration. Category-rating scales do not provide for that possibility. This disadvantage may well outweigh the benefits that the relatively effortless application of category-rating methods offer in comparison with magnitude estimation and multimodality matching techniques.[7]

NOTES

1. It is important to distinguish subjective "ratios" from mathematical ratios since it is empirically uncertain whether the numbers a subject produces in a magnitude estimation task preserve the respective subjective "ratios" or not. Ratios, therefore, should be distinguished from "ratios." This is done in the present text only, however, when the distinction is unclear otherwise.

2. Factor analyses are based on intercorrelations of 16 indexes that result from the complete tests of the measurement theoretical models. Following is the varimax rotated

factor matrix (normalized solution) for the 46 respondents to the occupational prestige study.

Indexes	Factor 1 ("Differences")	Factor 2 ("Ratios")
1. Quadruple Cond. Diff.	-.885	-.084
2. Independence Cond. Diff.	-.873	-.113
3. Stress Diff.	-.751	-.074
4. Compatibility Cat-09	.668	.468
5. Compatibility Cat-20	.728	.368
6. Linearity Cat-09	.795	.119
7. Linearity Cat-20	.811	.059
8. Quadruple Cond. Ratio	-.648	-.373
9. Independence Cond. Ratio	-.588	-.407
10. Stress Ratio	-.083	-.423
11. Compatibility Magnitude	.382	.73
12. Power Relation Magnitude	.355	.604
13. Deviation of ICMM-Function	.020	-.540
14. ICMM-Correlation	.059	.658
15. Fit Cat-09 to MAG	.239	.783
16. Fit Cat-20 to MAG	.161	.833
Factor contributions	5.505	3.811

In addition to the indexes mentioned thus far, the analysis included: (2) and (9) "independence," an auxiliary condition of algebraic measurement structures for testing the ordinal properties of the data (Orth, 1982a); (3) and (10) stress of one-dimensional MINISSA solutions of difference and ratio judgments; (13) deviation of the Indirect Cross-Modality Matching exponent from its predicted value (1.0 if numerical magnitude estimation and line production is involved); (14) correlation between magnitude modalities.

3. Version IV of the program was used since version V, with its alternative estimation procedures, was not available at the time of analyses.

4. Regression analyses included as additional independent variables the correlation between response modalities, the deviations of the direct and indirect cross-modality matching exponents from their expected, "theoretical" values, stress values of nonmetric analyses of the pair-estimation matrices, and also the percentages of violations of "independence" (see Note 2).

5. Contrary to the analyses of the psychophysical study, the analyses of the social judgments encountered some problems of convergence (Boomsma, 1982). Of the 46 best models, 15 did not reach convergence with the maximal number of iterations set at 1000. Seven of these yielded improper solutions. In general, however, the estimated coefficients do not diverge greatly from those of the other subjects, in spite of these deficiencies.

6. At this point it is informative to study how individual subjects differ in their judgment behavior in order to decide on strategies for the aggregation of scale values (Hannan, 1971). Thus there are significant but low positive correlations between the values of the λ-parameters with the age of respondents (e.g., .34 with regard to the prestige variable) and significant but low negative correlations with respondents' education (-.23 for prestige). Breaking down the 46 respondents of the study into 2 age and 3 educational

groups in a 2-by-3 design, however, and employing LISREL's multiple group option, it can be shown that the variability of the judgment functions (λ-parameters) is greater *between* the 6 resulting groups than *within* each group, while the parameters of the structural models (ψ) vary more *within* than *between* groups. These results, which cannot be dealt with in detail here, suggest that aggregation of scale values should be restricted to those respondents belonging to specific socioeconomic groups and who exhibit similar judgment functions.

7. It should be noted, however, that progress has been made in recent years toward an easily manageable implementation of bimodal magnitude techniques in interviewing (with number and line responses) such that, in comparison with category scaling, differences in terms of time and training have diminished greatly. For details of these procedures see Lodge (1981) and Wegener (1978, 1980, 1982b).

REFERENCES

ACOCK, A. C. and J. D. MARTIN (1974) "The undermeasurement controversy: should ordinal data be treated as interval?" Sociology and Social Research 63: 427-433.

ALLERBECK, K. R. (1978) "Messniveau und Analyseverfahren—Das Problem "strittiger Intervallskalen". Zeitschrift fuer Soziologie 7: 199-214.

BECK, U., M. BRATER and B. WEGENER (1979) Berufswahl und Berufszuweisung. Zur sozialen Verwandtschaft von Ausbildungsberufen. Frankfurt: Campus.

BENTLER, P. M. and D. G. BONETT (1980) "Significance tests and goodness of fit in the analysis of covariance structures." Psych. Bull. 88: 588-606.

BIRNBAUM, M. H. (1982) "Controversies in psychological measurement," in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

BLOCK, H. D. and J. MARSCHAK (1960) "Random orderings and stochastic theories of responses", in I. Olkin et al. (ed.) Contributions to Probability and Statistics. Stanford, CA: Stanford Univ. Press.

BOOMSMA, A. (1982) "The robustness of LISREL against small sample sizes in factor analysis models", in K. G. Joereskog and H. Wold (eds.) Systems under Indirect Observation: Causality, Structure, Prediction. Amsterdam: Elsevier North-Holland.

COXON, A.P.M. and C. L. JONES (1978) The Images of Occupational Prestige. A Study in Social Cognition. London: Macmillan.

CROSS, D. V. (1982) "On judgments of magnitude," in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

——— (1974) "Some technical notes on psychophysical scaling," in H. Moskowity et al. (eds.) Sensation and Measurement: Papers in Honor of S.S. Stevens. Dordrecht: Reidel.

DAVIS, K. and W. E. MOORE (1945) "Some principles of stratification." Amer. Soc. Rev. 10: 242-249.

HANNAN, M. T. (1971) Aggregation and Disaggregation in Sociology. Lexington, MA: D. C. Heath.

HAUSER, R. M. and A. S. GOLDBERGER (1971) "A treatment of unobservable variables in path analysis", in H.N. Costner (ed.) Sociological Methodology. San Francisco: Jossey-Bass.

HODGE, R. W., D. J. TREIMAN, and P. H. ROSSI (1966) "A comparative study of

occupational prestige," in R. Bendix and S. M. Lipset (eds.) Class, Status, and Power. New York: Free Press.

HODGE, R. W., P. M. SIEGEL, and P. H. ROSSI (1964) "Occupational prestige in the United States, 1925-1963." Amer. J. of Sociology 70: 286-302.

JOERESKOG, K. G. (1974) "Analyzing psychological data by structural analysis of covariance matrices," in D. H. Krantz et al. (eds.) Contemporary Developments in Mathematical Psychology, vol. 2.

——— (1971) "Statistical analysis of sets of congeneric tests." Psychometrika 36: 109-133.

——— and D. SOERBOM (1978) LISREL IV: A General Computer Program for Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods. Chicago: International Educational Services.

KRANTZ, D. H. (1972) "Magnitude estimation and cross-modality matching." J. of Mathematical Psychology 9: 168-199.

——— R. D. LUCE, P. SUPPES, and A. TVERSKY (1971) Foundations of Measurement, Vol. 1. New York: Academic.

LODGE, M. (1981) Magnitude Scaling. Quantitative Measurement of Opinion. Beverly Hills, CA: Sage.

LUCE, R. D. and W. EDWARDS (1958) "The derivation of subjective scales from just noticeable differences." Psych. Rev. 65: 222-237.

LUCE, R. D. and E. GALANTER (1963) "Psychophysical scaling", in R. D. Luce and E. Galanter (eds.) Handbook of Mathematical Psychology, vol. 1. New York: John Wiley.

MARKS, L. E. (1974) "On scales of sensation: prolegomena to any future psychophysics that will be able to come forth as a science." Perception & Psychophysics 16: 358-376.

ORTH, B. (1982a) "A theoretical and empirical study of scale properties of magnitude estimation and category rating scales," in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

——— (1982b) "Zur Bestimmung der Skalenqualitaet bei direkten Skalierungsverfahren." Zeitschrift fuer experimentelle und angewandte Psychologie 29: 160-178.

——— (1979) Rating-Verfahren und Groessenschaetz-Methode: Skalenniveau und funktionale Zusammenhaenge zwischen Skalen. Ph.D. dissertation, University of Kiel.

——— and B. WEGENER (1981) Einstellungsmessungen mit Magnitude- und Rating-Skalen-Modellen in einer Feldstudie. 23. Tagung experimentell arbeitender Psychologen, Berlin.

PARDUCCI, A. (1982) "Category ratings: still more contextual effects" in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

REISS, A. J. (1961) Occupations and Social Status. New York: Free Press.

SARIS, W. E. (1981) Different Questions, Different Variables? Unpublished manuscript, Amsterdam.

——— NEIJENS, P., and L. VAN DOORN (1980a) "Scaling social variables by multimodality matching." Methoden en Data Nieuwsbrief 5: 3-21.

——— (1980b) The Measurement of Occupational Prestige by Psychophysical Scaling. Unpublished manuscript, Amsterdam.

SCHNEIDER, B. (1982) "A nonmetric analysis of difference judgments in social psychophysics: scale validity and dimensionality," in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

——— S. PARKER, D. OSTROSKY, D. STEIN, and G. KANOW (1974) "A scale for the psychological magnitude of number." Perception & Psychophysics 16: 43-46.

SHEPARD, R. N. (1978) "On the status of 'direct' psychological measurement," in C. W. Savage (ed.) Minnesota Studies in the Philosophy of Science, vol. 9. Minneapolis: Univ. of Minnesota Press.

STEVENS, J. C. and M. GUIRAO (1964) "Individual loudness functions." J. of the Acoustical Society of America 36: 2210-2213.

STEVENS, S. S. (1975) Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. New York: John Wiley.

——— (1951) "Mathematics, measurement, and psychophysics", in S. S. Stevens (ed.) Handbook of Experimental Psychology. New York: Wiley.

——— (1946) "On the theory of scales of measurement." Science 103: 677-680.

——— and F. H. GALANTER (1957) "Ratio scales and category scales for a dozen perceptual continua." J. of Experimental Psychology 54: 377-411.

SUPPES, P. and J. L. ZINNES (1963) "Basic measurement theory," in R. D. Luce et al. (eds.) Handbook of Mathematical Psychology, vol. 1. New York: John Wiley.

TORGERSON, W. S. (1961) "Distances and ratios in psychological scaling." Acta Psychologica 19: 201-205.

TORGERSON, W. S. (1960) "Quantitative judgment scales," in H. Gulliksen and S. Messick (eds.) Psychological Scaling: Theory and Applications. New York: John Wiley.

TORGERSON, W. S. (1961) "Distances and ratios in psychological scaling." Acta Psychologica 19: 201-205.

TREIMAN, D. J. (1977) Occupational Prestige in Comparative Perspective. New York: Academic.

WEGENER, B. (1982a) "Outline of a structural taxonomy of sensory and social psychophysics," in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

——— (1982b) "Fitting category to magnitude scales for a dozen survey-assessed attitudes," in B. Wegener (ed.) Social Attitudes and Psychophysical Measurement. Hillsdale, NJ: Erlbaum.

——— (1980) "Magnitude-Messungen in Umfragen: Kontexteffekte und Methode." Zumanachrichten 6: 4-40.

——— (1979) "Magnitude-Messungen beruflicher Einstellungen," in U. Beck et al. Berufswahl und Berufszuweisung. Zur sozialen Verwandtschaft von Ausbildungsberufen. Frankfurt: Campus.

——— (1978) "Einstellungsmessung in Umfragen: Kategorische vs. Magnitude-Skalen." Zumanachrichten 3: 3-27.

——— and H. KIRSCHNER (1981) "A note on estimating interscale relations in 'direct' psychophysical scaling." British J. of Mathematical and Stat. Psychology 34: 194-204.

Bernd Wegener is Program Director at the Zentrum fuer Umfragen, Methoden und Analysen (ZUMA), Mannheim, West Germany. His research interests are with measurement issues and social perception. He has recently edited Social Attitudes and Psychophysical Measurement *(Erlbaum, 1982).*